HEINRICH-HERTZ-INSTITUT FÜR SCHWINGUNGSFORSCHUNG
BERLIN-CHARLOTTENBURG

Technischer Bericht Nr. 114

Sprachverarbeitung

von

Claus-Eberhard Liedtke

Berlin
1970

Technischer Bericht Nr. 114

Sprachverarbeitung

Zusammenfassung:

Für einen mittelgroßen Digitalrechner soll eine Sprachausgabe entworfen und simuliert werden. Der gesamte Wortvorrat soll im Kernspeicher des Rechners gespeichert sein. Das ist nur möglich, wenn die Redundanz der Sprache erheblich verringert wird. Zur Redundanzverringerung dienen Vocoder. Um den für diesen Verwendungszweck geeigneten Vocodertyp zu ermitteln, werden drei Konzepte durch Simulation auf dem Digitalrechner untersucht. Es handelt sich dabei um:

1. Kanalvocoder
2. Formantvocoder
3. Approximation der Sprach-Zeitfunktion durch Gaußfunktionen.

Es werden die Vocodertypen erläutert und Probleme, die bei ihrer Simulation auftreten, beschrieben.

Heinrich-Hertz-Institut für Schwingungsforschung

Der Bearbeiter

C.-E. Liedtke
(Liedtke)

Der Abteilungsleiter

(Prof. Dr.-Ing. W. Giloi)

Der Institutsdirektor

P Mathieu
(Prof. Dr. phil. P. Matthieu)

Berlin-Charlottenburg, den 9. März 1970

FORSCHUNGSBERICHT

ZUM THEMA

SPRACHVERARBEITUNG

von

Claus-Eberhard LIEDTKE

Heinrich-Hertz-Institut

Abteilung Informationsverarbeitung

Berlin, Februar 1970

Die Durchführung dieser Arbeit erfolgte mit Unterstützung der Deutschen Forschungsgemeinschaft

INHALT

Einleitung

Für den Digitalrechner CAE 90-40 des Lehrstuhls für Informationsverarbeitung soll eine Sprachausgabe entwickelt werden.
Bevor die hardwaremäßige Erstellung einer Sprachausgabevorrichtung in Angriff genommen werden kann, muß durch das Studium verschiedener Vocoderarten das optimale Konzept gefunden werden. Zu diesem Zweck werden verschiedene Möglichkeiten der Sprachkompression untersucht und auf dem Digitalrechner simuliert.
Die vollständige Simulation eines Sprachanalyse-Synthese-Systems, das auch Vocoder genannt wird, erfolgt in fünf Schritten:

1. Abtastung des analogen Sprachsignals
2. Analyse der Sprache nach wenigen charakteristischen Parametern
3. Zwischenspeicherung der Parameter
4. Synthese der Sprache aufgrund der gespeicherten Parameter auf dem Digitalrechner
5. Umformung der digitalen Information in ein analoges Sprachsignal, das hörbar gemacht werden kann.

Im folgenden soll nun die Durchführung der Schritte, wie sie im Rahmen dieser Arbeit bisher entworfen und realisiert wurden, ausführlich besprochen werden.

1. Analog-Digital- und Digital-Analog-Umsetzung

1.1 Signalvorverarbeitung

Die Verarbeitung von Sprache auf dem Digitalrechner kann nur dann durchgeführt werden, wenn die Information, hier die Sprache, in digitalisierter Form vorliegt. Das geschieht dadurch, daß man zunächst eine Spannung erzeugt, deren Zeitverlauf proportional dem Druck an einem Orte des Schallfeldes ist, sie geeignet vorverarbeitet und schließlich diese Spannung dann zu äquidistanten Zeitpunkten abtastet. Das Resultat ist eine Zahlenfolge, deren Glieder zeitlich fest zueinander in Beziehung stehen und deren Beträge der Amplitude des Drucks in einem Schallfeld entsprechen.

Eine Dynamikverringerung der Sprache ist insbesondere bei einer späteren Festkommarechnung auf dem Digitalrechner von Vorteil. Es wurde deshalb eine Dynamikkompression der Mikrophonspannung vorgesehen.

Durch die Abtastung des Analogsignals erfolgt eine Beschneidung des Frequenzbandes des Originalsignals. Nach dem Shannon'schen Abtasttheorem muß die Abtastfrequenz mindestens doppelt so groß wie die höchste zu verarbeitende Frequenz sein. Es wurde eine Abtastfrequenz von 10 kHz gewählt. Dieser Wert wird auch in der Literatur am häufigsten genannt.

Das Spektrum eines abgetasteten Vorgangs weist Überlagerungen durch die Faltung an den ganzen Vielfachen der Abtastfrequenz auf. Dadurch kann das ursprüngliche Spektrum nicht mehr aus der abgetasteten Zeitfunktion gewonnen werden.

Abb. 1 zeigt ein Beispiel für das Spektrum einer analogen Zeitfunktion, das der mit $f_T = 1/T$ abgetasteten Zeitfunktion und schließlich das Spektrum der Originalfunktion, wie es aus der abgetasteten Funktion berechnet werden würde. Man sieht, daß das berechnete Spektrum nicht mehr mit dem Spektrum der Originalfunktion übereinstimmt. Braucht man für die Weiterverarbeitung des Signals aber den Spektralverlauf des ursprünglichen Spektrums, muß man durch Tiefpaßfilterung des Signals dafür sorgen, daß die Überlagerungen vernachlässigbar klein werden. Abb. 2 zeigt die entsprechenden Spektren wie Abb. 1 für den Fall, daß die Originalfunktion mit einem Tiefpaß der Grenzfrequenz f_g gefiltert wurde. Man sieht, daß sich jetzt aus der abgetasteten Zeitfunktion das Spektrum der Originalfunktion weitaus genauer berechnen läßt, als im anderen Falle. Die Mikrophonspannung wurde deshalb durch einen Tiefpaß vorgefiltert. Bei dem Tiefpaß handelt es sich um ein

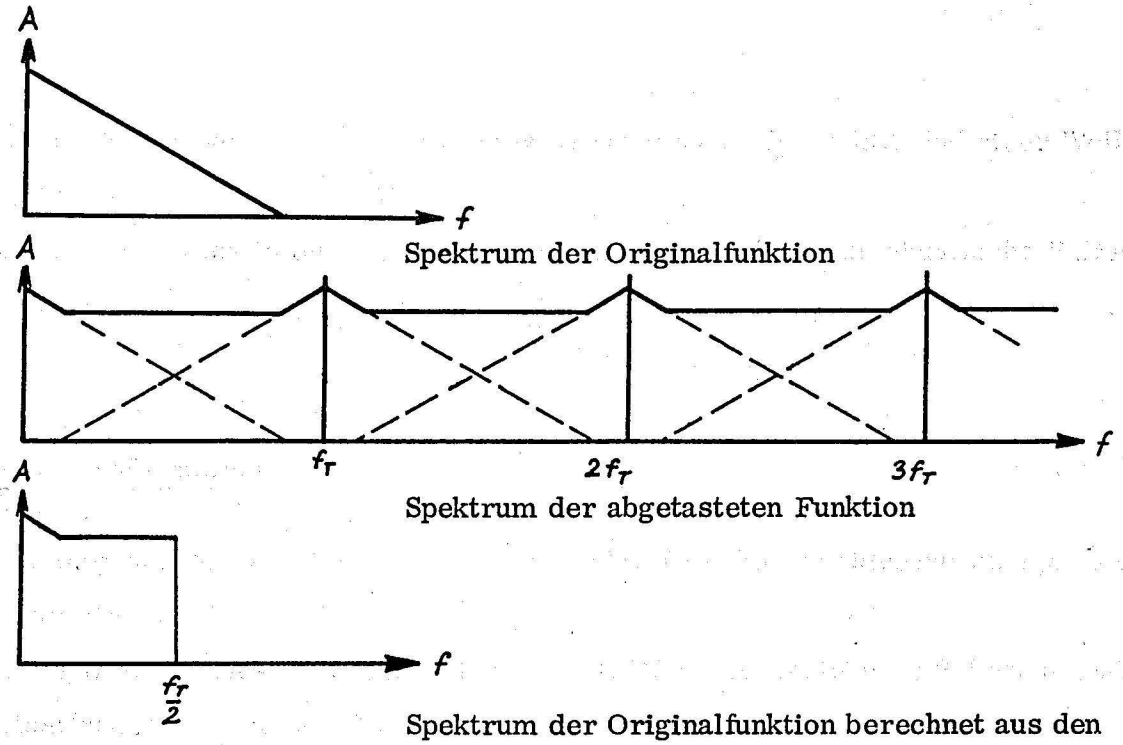
A
f
Spektrum der Originalfunktion
A
f_T
$2f_T$
$3f_T$
f
Spektrum der abgetasteten Funktion
A
f
$\frac{f_T}{2}$
Spektrum der Originalfunktion berechnet aus den Abtastwerten

Abb. 1

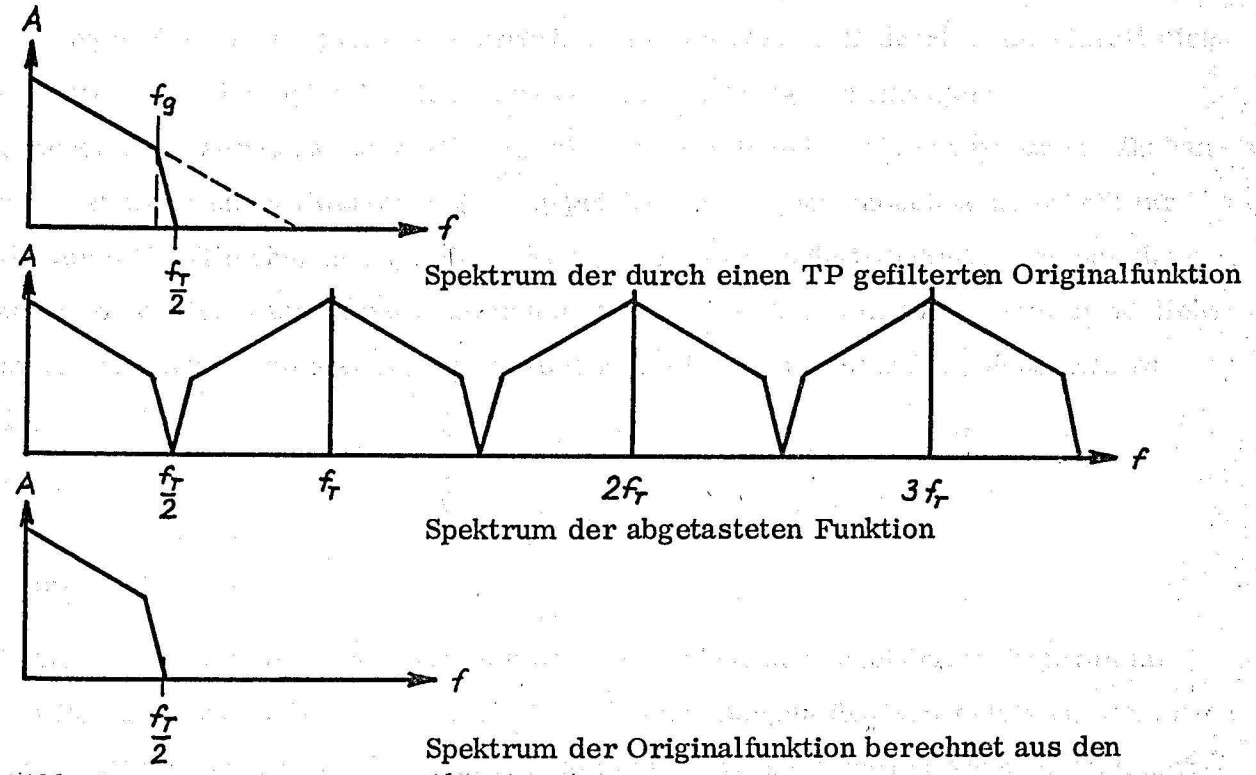
A
f_g
f
$\frac{f_T}{2}$
Spektrum der durch einen TP gefilterten Originalfunktion
A
$\frac{f_T}{2}$
f_T
$2f_T$
$3f_T$
f
Spektrum der abgetasteten Funktion
A
f
$\frac{f_T}{2}$
Spektrum der Originalfunktion berechnet aus den Abtastwerten

Abb. 2

dreipoliges Tschebyscheff-Filter mit einer Grenzfrequenz f_g = 4 kHz und einer Welligkeit von 1%.

Die Aufnahmen wurden in einem schalltoten Raum durchgeführt,um akustische Reflektionen auszuschalten.

1.2 Frequenztransformation

Die Abtastung des vorverarbeiteten Signals erfolgt über die Hybridrechenanlage CAE 90-40/ Telefunken RA 770.

Das vorliegende Hybridsystem mit seiner in FORTRAN geschriebenen Software gestattet nur Abtastfrequenzen bis 500 Hz.

Deshalb muß die analoge Zeitfunktion mindestens um den Faktor 20 in der Frequenz transformiert werden, wenn eine Tastfrequenz von 10kHz erreicht werden soll. Die Frequenztransformation wird mit einem Analog-Magnetbandgerät durchgeführt. Die analoge Zeitfunktion wird zunächst mit einer hohen Bandgeschwindigkeit aufgenommen und dann mit einer geringen Geschwindigkeit wiedergegeben und abgetastet. Dadurch multipliziert sich die Abtastfrequenz des Hybridsystems mit der Untersetzung des Bandgeräts.

Zur Untersetzung wurde das Magnetbandgerät 7001 von Brüel & Kjaer verwendet. Es handelt sich dabei um einen Zweikanal-Analogspeicher in Frequenzmodulationstechnik für Signale von 0 Hz (Gleichspannung) bis 20 kHz. Vier Bandgeschwindigkeiten ermöglichen Frequenz- bzw. Zeittransformationen in acht Stufen von 1 : 40 bis 40 : 1. Das ermöglicht zusammen mit der maximalen Abtastrate des Hybridsystems Abtastfrequenzen bis zu 20 kHz.

1.3 Abtastung

Die Analyse der Sprache kann wegen der Kompliziertheit der zugehörigen Programme nicht in Realzeit und zunächst auch nicht in der vierzigfachen Realzeit erfolgen. Deswegen würden die Abtastdaten schneller anfallen, als sie verarbeitet werden können. Man muß sich deshalb zunächst auf das Abtasten beschränken.

Da der Speicherplatz nur 16 K umfaßt, aber pro Sekunde abzutastender Sprache bereits 10000 Werte anfallen, müssen die Daten auf ein anderes externes Speichermedium transportiert werden und zwar zweckmäßigerweise auf Magnetband.
Um die Probleme, die bei der Abtastung auftreten, besser verstehen zu können, muß hier kurz etwas zum synchronisierten Datentransfer des Hybridsystems gesagt werden.
Der Hybridrechner arbeitet in zwei Modi, im Modus "CONTROL" und im Modus "WORK". Der Modus CONTROL entspricht der Analogrechnerstellung PAUSE und der Modus WORK der Analogrechnerstellung RECHNEN. Während des Modus WORK finden in aufeinanderfolgenden Zyklen hintereinander die folgenden Funktionsabläufe statt:

1. Datentransfer DR - AR
2. Datentransfer AR - DR
3. Rechnen (im Digitalrechner)
4. Warten bis Zyklusende.

Die Zeit eines solchen Zyklus wird auch Frametime genannt und kann per Programm eingestellt werden.
Da in jedem Zyklus der Datenaustausch zwischen Analog- und Digitalrechner nur einmal stattfindet, ergibt sich, daß die Frametime gleich dem Kehrwert der Abtastfrequenz sein muß. Bei einer gewünschten Abtastfrequenz von 10 kHz beträgt die Frametime 4 ms, wenn man eine vierzigfache Frequenztransformation durch das Analog-Magnetbandgerät berücksichtigt. In diesen 4 ms wird also ein Analogwert abgetastet und in den Kernspeicher des Digitalrechners gebracht. Die Zeit reicht aber nicht mehr dazu aus, den Abtastwert auf Magnetband zu schreiben. Der Abtastwert müßte in dem Fall als ein Block geschrieben werden. Der zu einem Block gehörende Anlauf- und Stoppschritt beträgt zusammen aber bereits 25 ÷ 30 ms. Die Abtastung muß deshalb so organisiert werden, daß zunächst ein Zahlenfeld von beispielsweise 5000 Zellen im Kernspeicher aufgefüllt und die Abtastung danach durch den Rücksprung vom Modus WORK in den Modus CONTROL unterbrochen wird. Im Modus CONTROL wird das Zahlenfeld auf Magnetband geschrieben, und anschliessend wird der nächste Analogblock im Modus WORK abgetastet. Ist die Abtastung eines Blocks beendet, muß auch das Analogmagnetbandgerät angehalten werden und erneut gestartet werden, wenn der nächste Block abgetastet werden soll.

1.4 Markierung

Es besteht die Gefahr, daß zwischen Starten und Stoppen des Analogmagnetbandgerätes ein Teil der Information verlorengeht. Deshalb werden Anfang und Ende eines jeden Analogblocks auf einer zweiten Spur des Magnetbands durch einen Puls markiert. Die Markierung auf dem Analog-Magnetband kann durch Betätigung einer Taste von Hand erfolgen oder vom Rechner durch ein Markierprogramm. Es ist dabei lediglich zu beachten, daß die Länge eines Analogblocks nicht den Wert überschreitet, der 5000 Abtastwerten entspricht, da nicht mehr als 5000 Zellen im Kernspeicher für die Abtastwerte reserviert worden sind.

Wird das Magnetbandgerät gestartet und befindet sich der Hybridrechner im Modus WORK, so leitet der erste Markierpuls über eine Interruptleitung den Abtastvorgang ein, während der zweite Markierpuls über eine zweite Interruptleitung die Abtastung beendet. Das Bandgerät wird selbsttätig angehalten und muß um einige Zentimeter zurückgespult werden, damit es beim erneuten Start die letzte Endmarke jetzt als Anfangsmarke lesen kann.

1.5 Digital-Analog-Umsetzung

Bei der Digital-Analog-Umsetzung, die aus den Abtastwerten synthetischer Sprache wieder eine kontinuierliche Zeitfunktion auf dem Analog-Magnetband erzeugen soll, tritt ein ähnliches Problem wie bei der Abtastung auf. Die Frametime, die jetzt wiederum 4 ms beträgt, reicht nicht aus, um den Wert vom Digital-Magnetband zu lesen. Deshalb wird entsprechend zur Abtastung in Modus CONTROL ein Block im Kernspeicher mit den synthetisch erzeugten Abtastwerten gefüllt. Im Modus WORK wird durch die Anfangsmarke auf dem Analog-Magnetband ein Interrupt erzeugt, der die Ausgabe der Analogwerte einleitet. Am Ende eines ausgegebenen Blocks wird auf die zweite Spur zusätzlich eine Marke gesetzt, die das Ende des Analogblocks auf Spur 1 anzeigt. Im Modus CONTROL wird dann der nächste Block vom Digital-Magnetband gelesen und das Analog-Magnetband um einige Zentimeter zurückgespult, so daß bei einer erneuten Ausgabe die vom Rechner unmittelbar vorher geschriebene Endmarke wieder als Anfangsmarke gelesen werden kann.

2. Kanal- und Formantvocoder

2.1 Natürliche Spracherzeugung und Simulation

Die menschlichen Sprachlaute kann man ganz grob in stimmhafte und stimmlose Laute unterteilen. Die "Quelle" für die stimmhaften Laute liegt an der Stimmritze. Diese sendet mit quasikonstanter Frequenz Luftpulse in den Vokaltrakt.
Stimmlose Laute werden auf verschiedene Art und Weise erzeugt. Die "Quelle" kann in diesem Falle aus einer Konstriktion im Vokaltrakt bestehen, durch die Luft gepreßt wird. Die "Quelle" gibt ein Zischen ab. Eine andere Möglichkeit für die Entstehung stimmloser Laute ist die, daß der Trakt zunächst an einer Stelle geschlossen wird und sich hinter dieser ein Druck aufbaut. Der Druck entweicht plötzlich, wenn die Stelle des Traktes kurzzeitig geöffnet wird.
Die "Quelle" regt den Vokaltrakt zu Schwingungen an.
Der Vokaltrakt besteht im wesentlichen aus den Mund-, Nasen- und Rachenhöhlen. Die Übertragungseigenschaften des Traktes lassen sich durch Bewegung des Unterkiefers, der Zunge, der Lippen und des Velums variieren.
Mathematisch läßt sich die Spracherzeugung durch die folgende Gleichung beschreiben:
Im Frequenzbereich gilt: $P(s) = S(s) \cdot T(s)$ (1)
bzw. im Zeitbereich gilt: $p(t) = s(t) * t(t)$ (2)
In Gl. (1) bedeuten

1.) P(s) das Spektrum des Drucks im Schallfeld eines sprechenden Menschen
2.) S(s) das Spektrum der Quelle
3.) T(s) die Übertragungsfunktion des Vokaltraktes.

Die Faktoren in Gl.(2) stellen die Zeitverläufe bzw. Impulsantworten dar,und der Stern deutet die Faltung an.
Will man künstliche Sprache erzeugen, so kann man dies dadurch erreichen, daß das menschliche Spracherzeugungssystem simuliert wird.
Das bedeutet folgendes: Man muß sich zunächst eine Quelle erzeugen, die entweder Pulse zur Erzeugung stimmhafter Laute oder Rauschen für stimmlose Laute liefern kann. Zur Steuerung dieser sehr einfachen Quelle sind zwei Parameter notwendig. Der erste, hier

LQF genannt, gibt die sog. Pitchfrequenz an, das ist die Frequenz des Pulsgenerators und der zweite Parameter, der hier LVU heißt, bestimmt, zu welchem Zeitpunkt die Quelle Rauschen und zu welchem sie Pulse abgeben soll.
Die Quelle wird auf ein Filter geschaltet, das in jedem Augenblick die Übertragungseigenschaften des Vokaltraktes aufweist.
Aus dem gesagten ergibt sich, daß folgende Schritte zur Erzeugung künstlicher Sprache notwendig sind:
Im ersten Schritt wird "echte" Sprache auf ihre Eigenschaften untersucht. Die Ergebnisse dieser Analyse sind die Steuerparameter der Quelle und die Übertragungseigenschaften des Vokaltraktes.
Im zweiten Schritt werden diese Angaben für die Synthese künstlicher Sprache verwendet.

2.2 Syntheseteil des Kanalvocoders

Sowohl beim Kanalvocoder wie beim Formantvocoder wird ein Filter benötigt, das in jedem Augenblick den Frequenzgang des Vokaltraktes nachbildet. Die Phaseninformation der Übertragungsfunktion T(s) wird dabei vernachlässigt. Das ist zulässig, da Versuche die relativ geringe Bedeutung der Phaseninformation für die Verständlichkeit von Sprache nachgewiesen haben. Ein typisches Beispiel für den Frequenzgang des Vokaltraktes stellt G_{dB1} in Abb. 3 dar.
Beim Kanalvocoder wird der Frequenzbereich des Vokaltraktes durch eine sog. Filterbank in n Abschnitte unterteilt. Die Filterbank besteht aus n parallelgeschalteten Bandpässen. Die oberen und unteren Grenzen der benachbarten Bandpässe wurden gerade so gewählt, daß sie sich in ihrem 3 dB-Abfall überschneiden. Ein Beispiel für die Übertragungsfunktion einer Filterbank, die aus 10 Bandpässen besteht, ist in Abb. 4 dargestellt. Das Blockschaltbild eines Kanalvocoders zeigt Abb. 5. Wird auf die Eingänge sämtlicher Bandpässe (in Abb. 5 rechts vom Übertragungskanal) ein Puls gegeben, so schwingen alle Bandpässe mit gleicher Amplitude aber unterschiedlicher Frequenz, denn jeder Bandpaß schwingt mit seiner Mittenfrequenz. Würde man die Ausgangssignale B1 bis Bn aufsummieren, erhielte man eine Zeitfunktion, die sämtliche (Mitten-) Frenquenzen gleichmäßig überträgt.

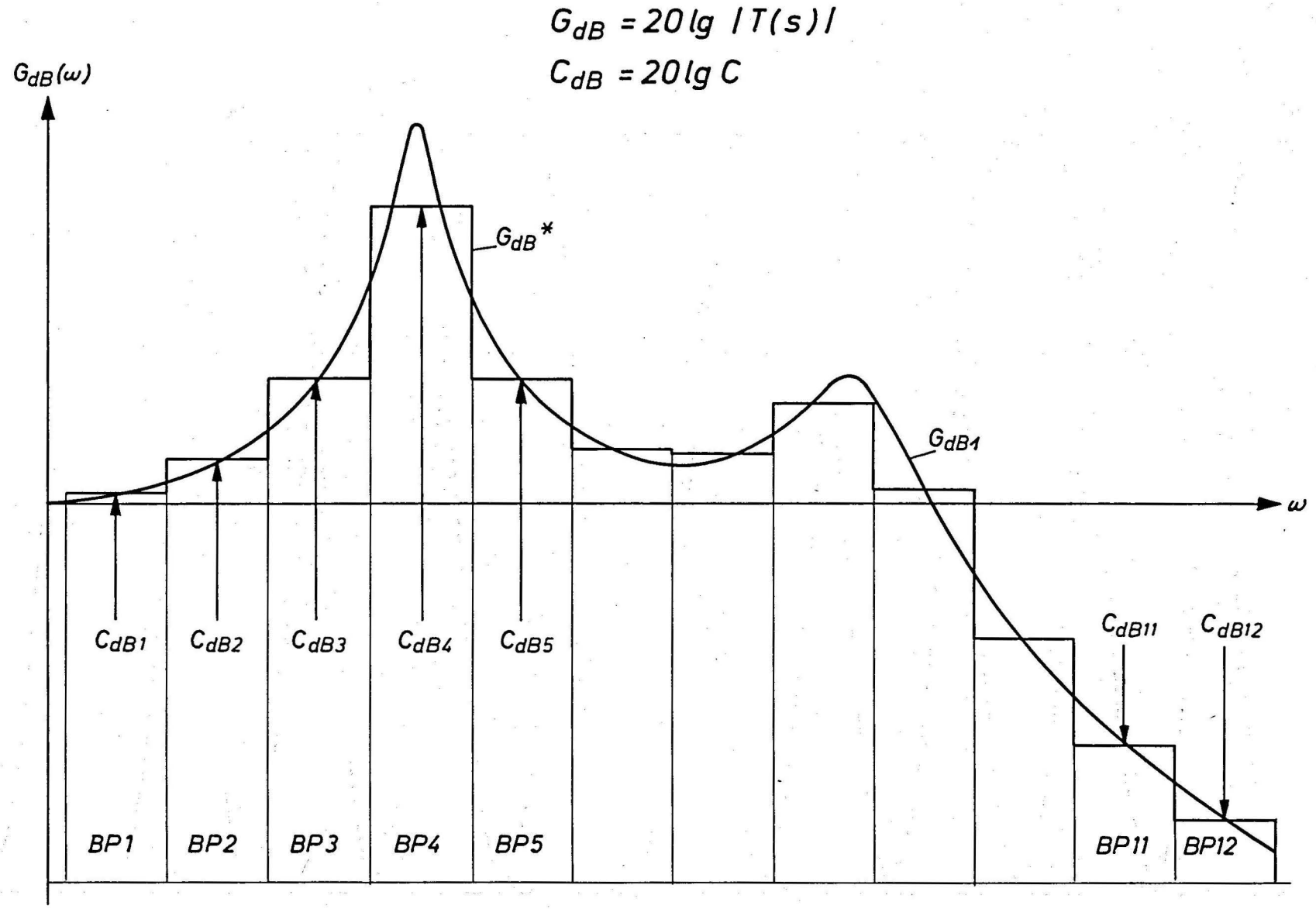

Abb.3 Approximation des Frequenzganges durch eine Treppenkurve beim Kanalvocoder.

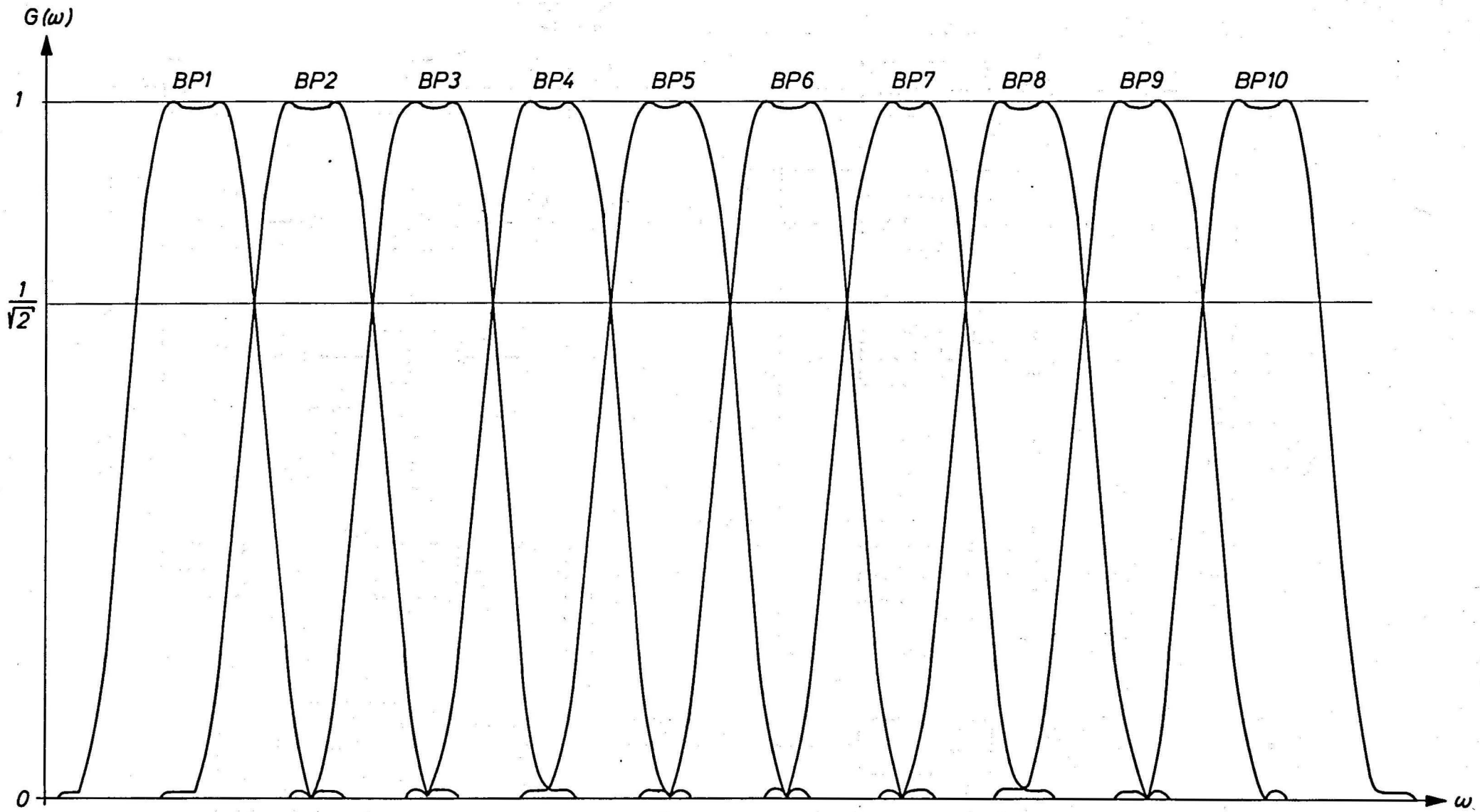

Abb. 4 Beispiel für Übertragungsfunktionen einer Filterbank aus 10 Bandpässen.

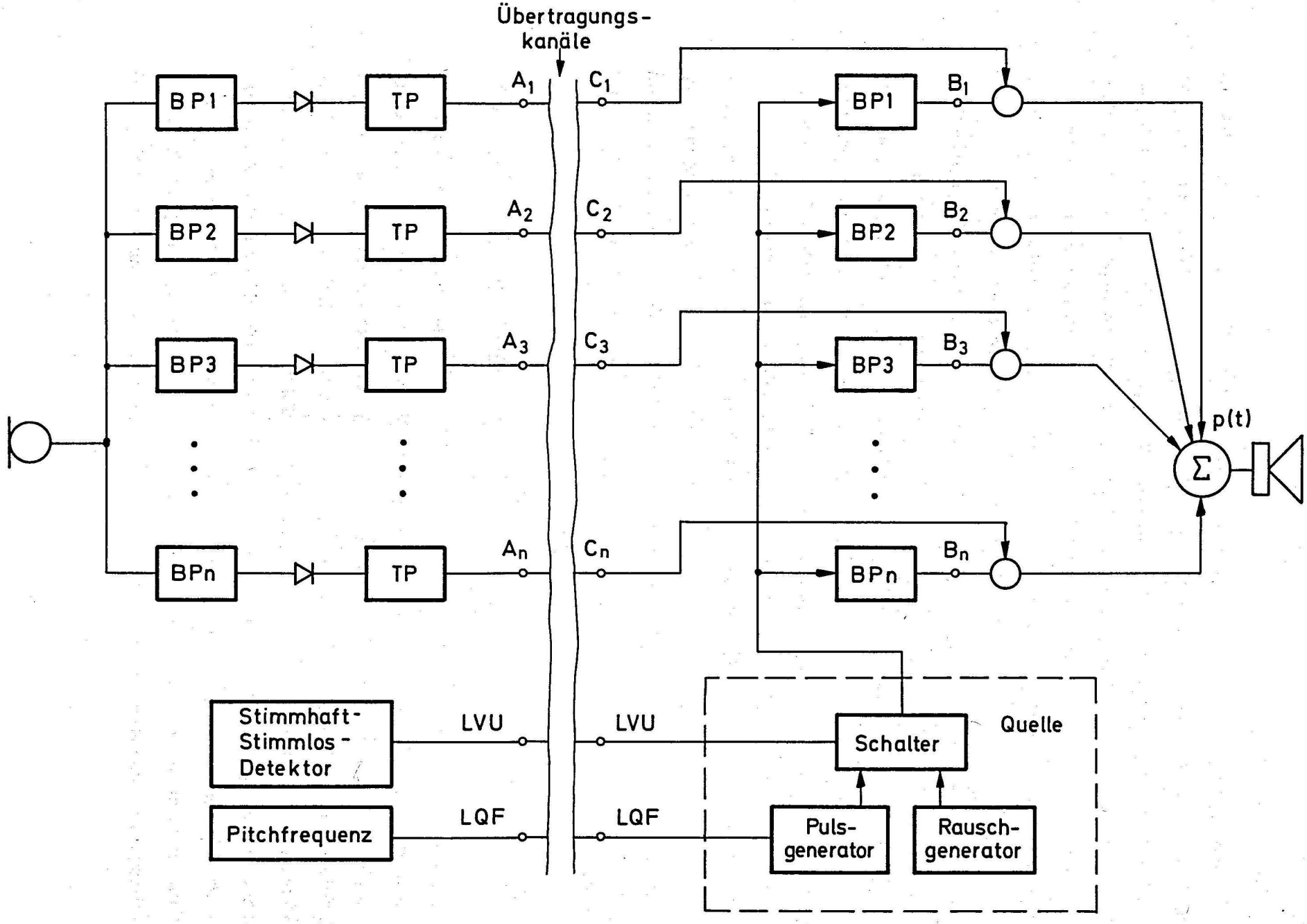

Abb. 5 Kanalvocoder

Die Filterbank weist die Übertragungsfunktion eines Allpasses auf. Bewertet man die Ausgänge der Bandpässe z. B. durch Koeffizienten c_i, deren Größe man aus Abb. 3 im logarithmischen Maßstab als c_{dBi} entnehmen kann, dann werden im Summensignal p(t) nicht mehr alle Frequenzen gleich stark vertreten sein, sondern die einzelnen Frequenzen werden je nach Bewertung durch die c_i erscheinen. Damit ist die Filterbank kein Allpaß mehr, sondern sie stellt ein Filter dar, dessen Frequenzgang sich durch die Koeffizienten c_i als Treppenkurve G_{dB}^* (siehe Abb. 3) einstellen läßt. So kann man jeden Frequenzgang approximieren. Die Approximation wird umso besser sein, je mehr Bandpässe vorliegen, d.h. auch je schmalbandiger die Filter werden.
Ein Kanalvocoder dient dazu, bei der Übertragung eines Sprachsignales die Redundanz der Sprache zu verringern. Die Information der Sprache liegt außer in den Parametern - LQF - der Pitchfrequenz und - LVU -, der Stimmhaft- Stimmlos-Entscheidung für die Quelle in den Koeffizienten c_i des Filters. Will man eine hohe Sprachkompression erreichen, d.h. nur wenige Bit zur Charakterisierung der Sprache verwenden, darf man auch nur wenige Kanäle, d.h. Bandpässe, verwenden.
Im Zusammenhang mit der vorliegenden Arbeit wurde ein Kanalvocoder auf dem Digitalrechner simuliert, der aus 15 Kanälen besteht.

2.3 Analyseteil des Kanalvocoders

Mit der oben beschriebenen Syntheseschaltung läßt sich ein Filter aufbauen, das in jedem Augenblick den Frequenzgang des Vokaltraktes nachbilden kann unter der Voraussetzung, daß auch für jeden Augenblick die Parameterkombination c_i bekannt ist. Die Analyseschaltung ist in Abb. 5 links von den Übertragungskanälen zu sehen. Die Sprachzeitfunktion, die z. B. durch Abtastung einer Mikrophonspannung gewonnen werden kann, wird auf eine Filterbank gegeben. Die Filterbank ist genauso aufgebaut wie die, die bereits in der Syntheseschaltung beschrieben wurde. Die Ausgänge der Bandpässe werden gleichgerichtet und durch Tiefpässe geglättet. Die Ausgangssignale A_i sind dann zu jedem Augenblick dem Betrag des Spektrums im Durchlaßbereich des entsprechenden Bandpasses proportional.

Da die Bandpässe aber den ganzen Frequenzbereich lückenlos überstreichen, stellen die Ausgänge A_i die Approximation des Spektrums durch eine Treppenkurve dar, wie es in Abb. 3 als G_{dB}^{*} dargestellt wurde. Daraus ergibt sich, daß die A_i gerade die Koeffizienten sind, die als Parameter c_i zur Steuerung des Syntheseteils benötigt werden.

2.4 Darstellung des Vokaltrakts durch Formanten

Die Übertragungsfunktion des Vokaltraktes, T(s), ist eine reelle Funktion und läßt sich daher durch konjugiert komplexe Nullstellen- und Polpaare bzw. einfach reelle Nullstellen und Pole beschreiben. Wird die Funktion so normiert, daß $T(0) = 1$ ergibt, erhält man die Schreibweise

$$T(s) = \prod_{i=1}^{n} \frac{s_{pi} \cdot s_{pi}^{*}}{(s-s_{pi})(s-s_{pi}^{*})} \cdot \prod_{j=1}^{m} \frac{(s-s_{zj})(s-s_{zj}^{*})}{s_{zj} \cdot s_{zj}^{*}} \qquad (3)$$

Ein Polpaar mit der Übertragungsfunktion

$$G(s) = \frac{s_p \cdot s_p^{*}}{(s-s_p)(s-s_p^{*})} \qquad \text{wobei} \quad \begin{aligned} s_p &= -\sigma_p + j\omega_p \\ s_p^{*} &= -\sigma_p - j\omega_p \end{aligned} \qquad (4)$$

wird als Formant und ein Nullstellenpaar mit der Übertragungsfunktion

$$H(s) = \frac{(s-s_z)(s-s_z^{*})}{s_z \cdot s_z^{*}} \qquad \text{wobei} \quad \begin{aligned} s_z &= -\sigma_z + j\omega_z \\ s_z^{*} &= -\sigma_z - j\omega_z \end{aligned} \qquad (5)$$

ist, wird als Antiformant bezeichnet. Die Übertragungsfunktion nach Gl. (3) läßt sich als Produkt von Formanten und Antiformanten deuten. Die logarithmische Darstellung, die nach der Gl.

$$T_{dB}(\omega) = 20 \cdot \lg |T(s)| \qquad (6)$$

eingeführt wird, führt zur Gl.

$$T_{dB}(\omega) = \sum_{i=1}^{n} G_{dB}(\omega) + \sum_{j=1}^{m} H_{dB}(\omega) \qquad (7)$$

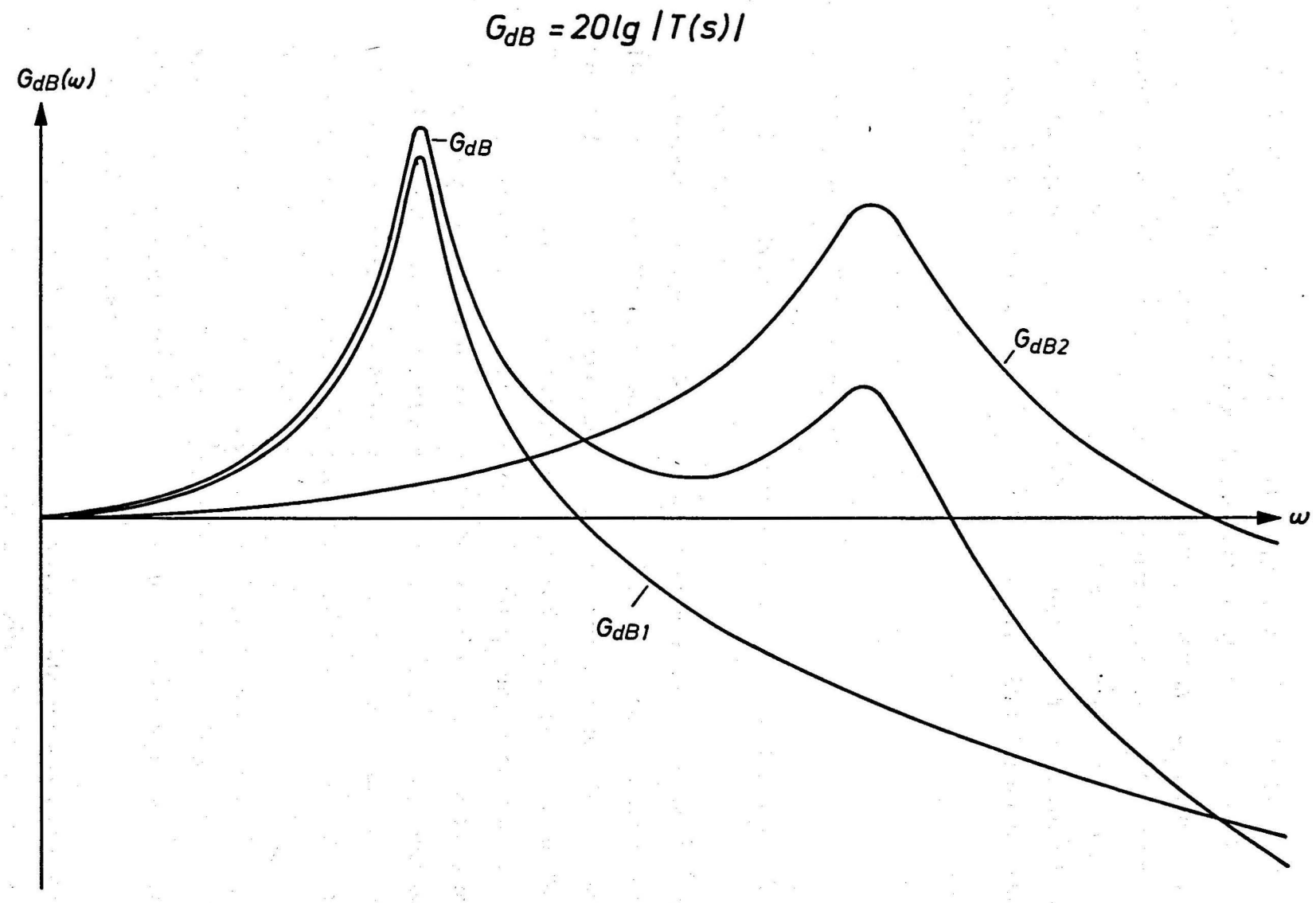

Abb. 6 Approximation des Frequenzganges G_{dB} durch zwei Formanten mit den Frequenzgängen G_{dB1} und G_{dB2}.

Abb. 6 zeigt, wie sich der Frequenzgang $G_{dB}(\omega)$ leicht durch die beiden Formanten $G_{dB1}(\omega)$ und $G_{dB2}(\omega)$ darstellen läßt. Es hat sich gezeigt, daß sich der Frequenzgang des Vokaltraktes bei nichtnasalen, stimmhaften Lauten gut durch die ersten drei Formanten beschreiben läßt. Da jeder Formant durch zwei Parameter, nämlich die Dämpfung σ_p und die Polfrequenz ω_p charakterisiert ist, braucht man zur Approximation der Übertragungsfunktion des Vokaltraktes weitaus weniger Parameter (im Beispiel 6 Parameter gegenüber den 15 beim Kanalvocoder) und erreicht dadurch eine höhere Sprachkompression.

2.5 Aufbau des simulierten Formantvocoders

Es wurde ein vollständiger Formantvocoder simuliert. Das Blockschaltbild der Syntheseschaltung zeigt Abb. 7.

Die Quelle des Formantvocoders besteht aus einem Rauschgenerator mit nachgeschaltetem Tiefpaß und einem Pulsgenerator mit nachgeschaltetem Pulse- shaping- network.

Der Tiefpaß begrenzt die Rauschquelle auf den interessierenden Bereich bis ca. 3500 Hz.

Das Pulse- shaping- network formt aus den Diracstößen des Pulsgenerators einen dreieckförmigen Verlauf, der etwa dem Zeitverlauf der Luftpulse an der Glottis entspricht.

Als Pulse- shaping- network wird ein Polpaar verwendet, bei dem die Dämpfung gerade doppelt so groß wie die Polfrequenz ist. Das entspricht gerade dem aperiodischen Grenzfall. Die Polfrequenz wird proportional zur Pitchfrequenz variiert. Die beiden Potentiometer mit den Parameterbezeichnungen MFRI und MVOC regulieren die Amplituden der beiden Generatorzweige.

Mit dem Schalter kann entweder der Rauschgeneratorzweig oder der Pulsgeneratorzweig auf das sich anschließende Formantnetzwerk geschaltet werden.

FLANAGAN berechnete einen Vokaltrakt für die Vokalerzeugung nichtnasaler Laute unter den folgenden Voraussetzungen.

1.) Der Vokaltrakt hat über seine ganze Länge l einen konstanten Querschnitt.

2.) Er wird in Richtung Mundöffnung durch den Strahlungswiderstand belastet, der klein gegen die Impedanz des Traktes ist.

3.) Der Trakt wird von der Glottis durch einen "Schnelle-Generator" erregt, dessen innerer Widerstand groß gegen den Eingangswiderstand des Traktes ist.

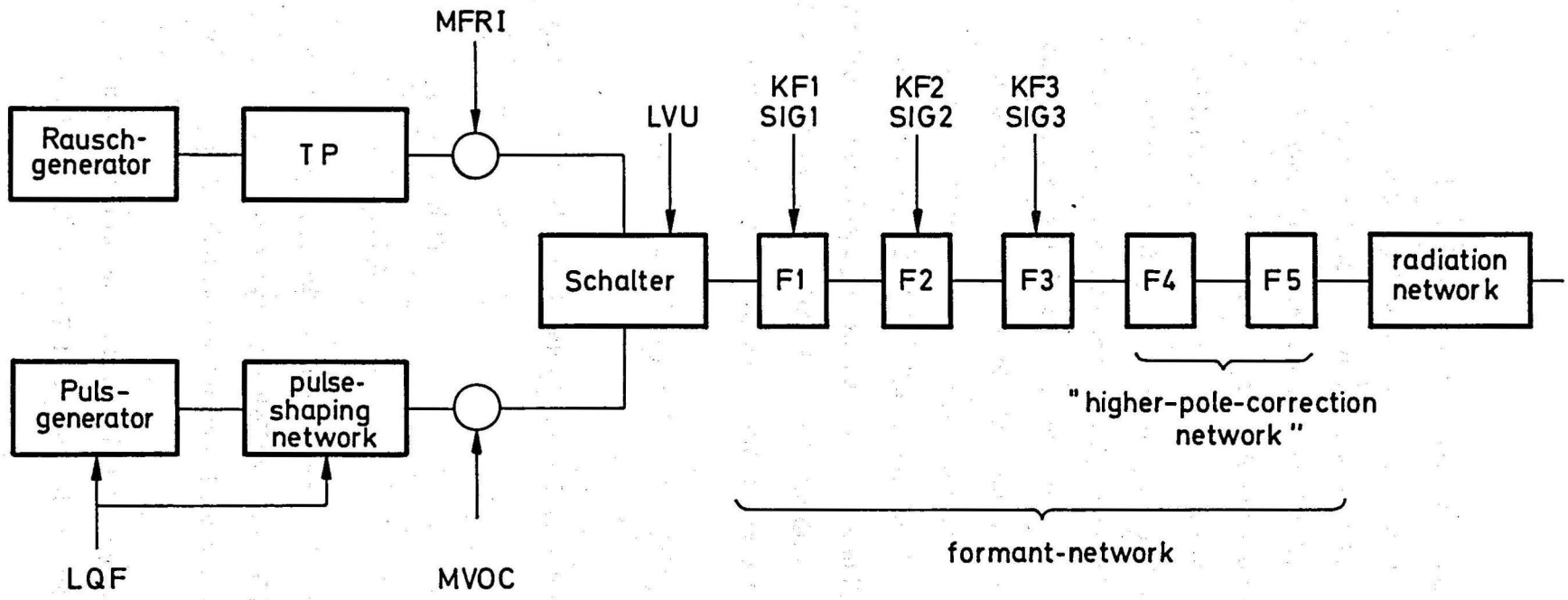

Abb. 7 Blockschaltbild für den Synteseteil des simulierten Formantvocoders.

Die mathematische Behandlung des verlustbehafteten Vokaltraktes führte zu dem Ergebnis, daß die Übertragungsfunktion des Vokaltraktes durch eine unendliche Anzahl konjugiert komplexer Polpaare charakterisiert wird[1]. Es treten keine Nullstellen auf.
Bei dem simulierten Formantvocoder sollen aber nur die ersten drei Polpaare F1, F2 und F3 verändert werden können. Es muß aber außer den drei ersten Formanten eine Korrektur für die fehlenden Polpaare durchgeführt werden. Bei einem digitalen Formantvocoder ist die Korrektur besonders einfach, da automatisch zu den eingeführten Formanten alle zur halben Samplingfrequenz und deren Vielfachen spiegelbildlich liegenden Formanten ebenfalls enthalten sind. Es braucht deshalb nur ein vierter und fünfter Formant mit den Polfrequenzen von 3500 Hz und 4500 Hz zur "higher- pole- correction" hinzugeführt zu werden. Die zugehörigen Bandbreitewerte von 175 Hz und 281 Hz wurden für eine Samplingfrequenz von 10 kHz von L. R. RABINER übernommen.
Die Berechnung des Schallfeldspektrums nach Gl. (1) enthält nur die beiden wichtigsten Glieder, die Eigenschaften der Quelle, S(s), und die des Traktes T(s). Will man die Abstrahlung des Schalls noch berücksichtigen, dann muß das Glied R(s) hinzugefügt werden.

$$P(s) = S(s) \cdot T(s) \cdot R(s) \qquad (8)$$

Approximiert man die Abstrahlung durch die Strahlung einer schwingenden Kugel, so ergibt sich als normalisierte Impedanz

$$z = \frac{jka}{1+jka} \qquad k = \frac{\omega}{c} \qquad (9)\ ^{2)}$$

Gl. (9) läßt sich in der komplexen Ebene durch einen Pol auf dem negativen Teil der reellen Achse und eine Nullstelle im Ursprung darstellen. Als Poldämpfung wurde $\sigma_p = 3500$ Hz gewählt. Das Filter, das die Abstrahlung berücksichtigt, wird in Abb. 7 mit Radiation-Network bezeichnet.
Das gesamte Syntheseprogramm ist flexibel aufgebaut und läßt sich durch den Einbau von Nullstellen, weiteren Formanten und Filtern beliebig erweitern.

1) FLANAGAN, S. 177, Gl. 6.5.
2) FLANAGAN, S. 33, Gl. 3.37.

Zur Steuerung des vorliegenden Formantvocoders sind

4 Parameter für die Quelle

und 6 Parameter für das Formantnetzwerk

notwendig. Damit ermöglicht der Formantvocoder eine höhere Sprachkompression als der Kanalvocoder.

Zum gegenwärtigen Zeitpunkt bestehen noch erhebliche Schwierigkeiten darin, die Steuerparameter in ökonomischer Weise zu bestimmen. Die Arbeiten werden auf diesem Gebiet deshalb besonders intensiv vorangetrieben.

3. Sprachentwicklung nach Gaußschen Funktionen

3.1 Prinzip

Im folgenden soll eine völlig andere Möglichkeit der Sprachanalyse-Synthese beschrieben werden, die sich aus der Betrachtung des Zeitverlaufs einer Sprachschwingung ergibt. Der Zeitverlauf möge in Segmente endlicher Länge zerlegt werden. Man kann dann die eindeutige Zuordnung der Spannungswerte eines Mikrophons zu den Zeitwerten mathematisch als Funktion f(t) auffassen und versuchen, diese durch einen Satz anderer Funktionen g anzunähern, so daß sich

$$f(t) = \sum_{i=1}^{n} a_i g_i (t - t_i) \qquad (10)$$

ergibt. Die Güte der Approximation wird beeinflußt durch günstige Segmentierung und die Wahl geeigneter Funktionen g_i.

Zunächst ein Wort zu "Wahl geeigneter Funktionen". Eine besonders rasche Approximation ist dann zu erwarten, wenn die Funktionen g_i bereits von sich aus einen ähnlichen Verlauf wie eine Sprachschwingung haben. Sie müssen in dem betrachteten Bereich dem Betrag nach endlich bleiben, mehrere Nullstellen aufweisen und für große Argumente gegen Null konvergieren. Schließlich, wenn man an eine hardwaremäßige Synthese denkt, soll sich

ihr Zeitverlauf möglichst einfach durch eine Rechenschaltung auf dem Analogrechner darstellen lassen. Diesen Anforderungen genügen beispielsweise die Gaußschen Funktionen.

Die Anregung zur Wahl der Gaußfunktionen geht zurück auf eine Veröffentlichung von J. A. HOWARD und R. C. WOOD. In der vorliegenden Arbeit wurde ein von Howard und Wood unabhängiger Weg für die Darstellung stimmhafter Laute durch Gaußsche Funktionen beschritten, der im folgenden beschrieben werden soll.

Die Gaußschen Funktionen n-ter Ordnung G_n sind die Lösungen der Differentialgleichung

$$G_n'' + t \cdot G_n' + (n+1) \cdot G_n = 0 . \qquad (11)$$

Für das Analyseprogramm wurden Gaußfunktionen bis zur 10. Ordnung vorgesehen. Die Praxis zeigt aber, daß nur Gaußfunktionen bis zur 5. Ordnung benötigt werden. Um eine Vorstellung vom Verlauf des Graphen der Funktionen zu geben, sind in Abb. 8, Abb. 9 und Abb. 10 Gaußfunktionen von nullter bis achter Ordnung graphisch dargestellt worden. Dabei ist zu beachten, daß die Gaußfunktionen gerader Ordnung sich als gerade Funktionen und die ungerader Ordnung sich als ungerade Funktionen in den Bereich negativer Argumente fortsetzen.

3.2 Segmentierung

Der erste Schritt in der Analyse ist die Aufspaltung der Zeitfunktion in die Bereiche, in denen die eigentliche Analyse stattfinden soll. Die Peak-Struktur stimmhafter Laute kommt der Segmentierung sehr entgegen, wenn man als Länge eines Segmentes gerade eine Pitch-Periode der Sprache nimmt. Es wurde deshalb zunächst eine Markierung der Sprache nach D. R. REDDY vorgenommen. Der Abstand zweier aufeinanderfolgender "Significant Maximum Peaks", die im folgenden mit SMP abgekürzt werden sollen, entspricht ja gerade einer Pitch-Periode. Der Bereich von einem SMP zum nächsten eignet sich jedoch nicht zur Entwicklung nach Gaußfunktionen, da die Gaußfunktionen gerade und ungerade Funktionen sind, die zu positiven und negativen Argumenten abfallen. Man muß deshalb auch den Analysebereich so wählen, daß das absolute Maximum, das hier durch den SMP gekennzeichnet ist, ebenfalls mehr in die Mitte des Analysebereichs verlegt wird. Die genaue Bestimmung des Analyseintervalls geschieht folgendermaßen: Es wird vom SMP zu positiven (bzw. negativen) Argumenten das Maximum M gesucht, das kleiner als das folgende Maximum M^+ ist.

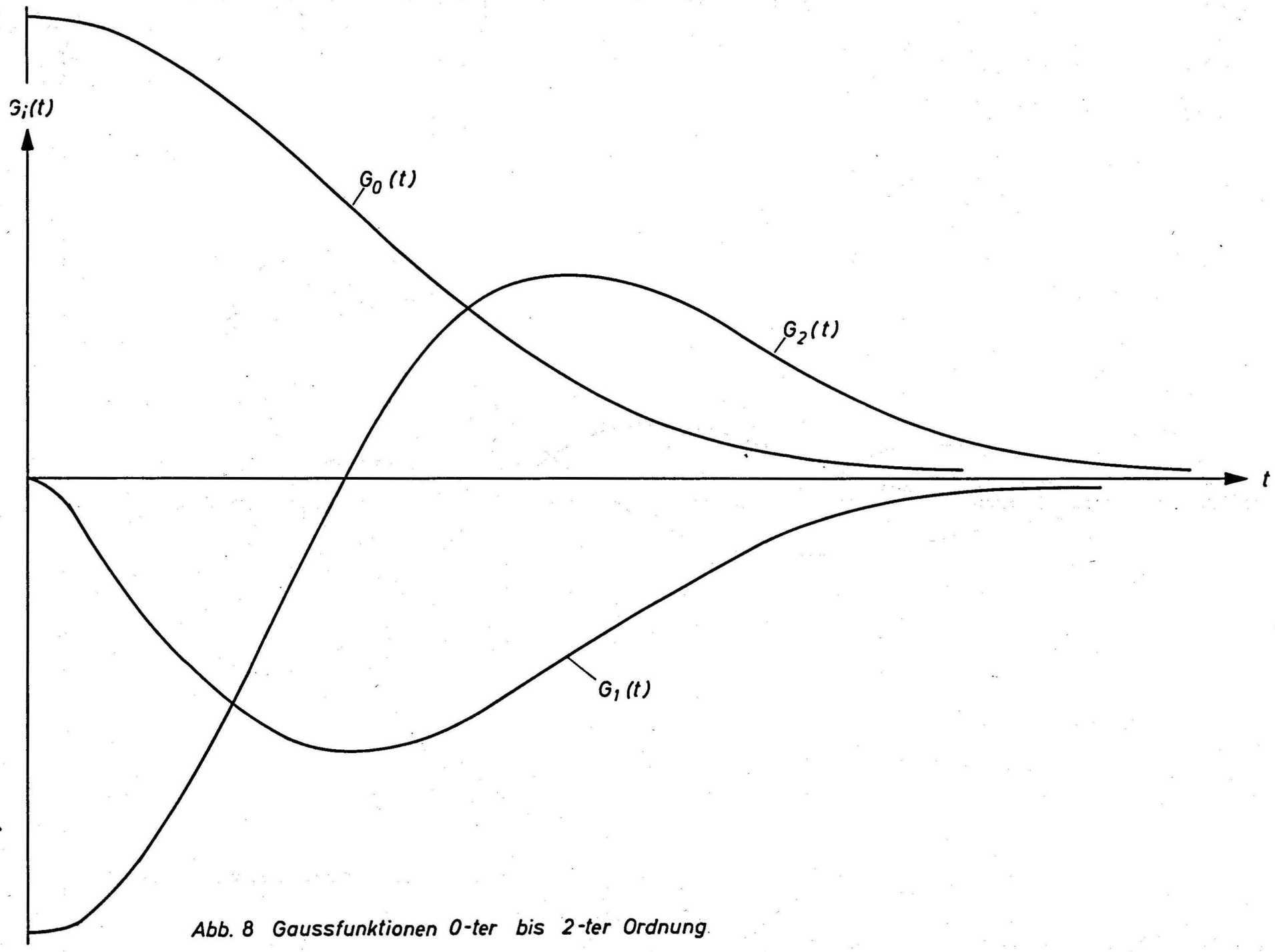

Abb. 8 Gaussfunktionen 0-ter bis 2-ter Ordnung

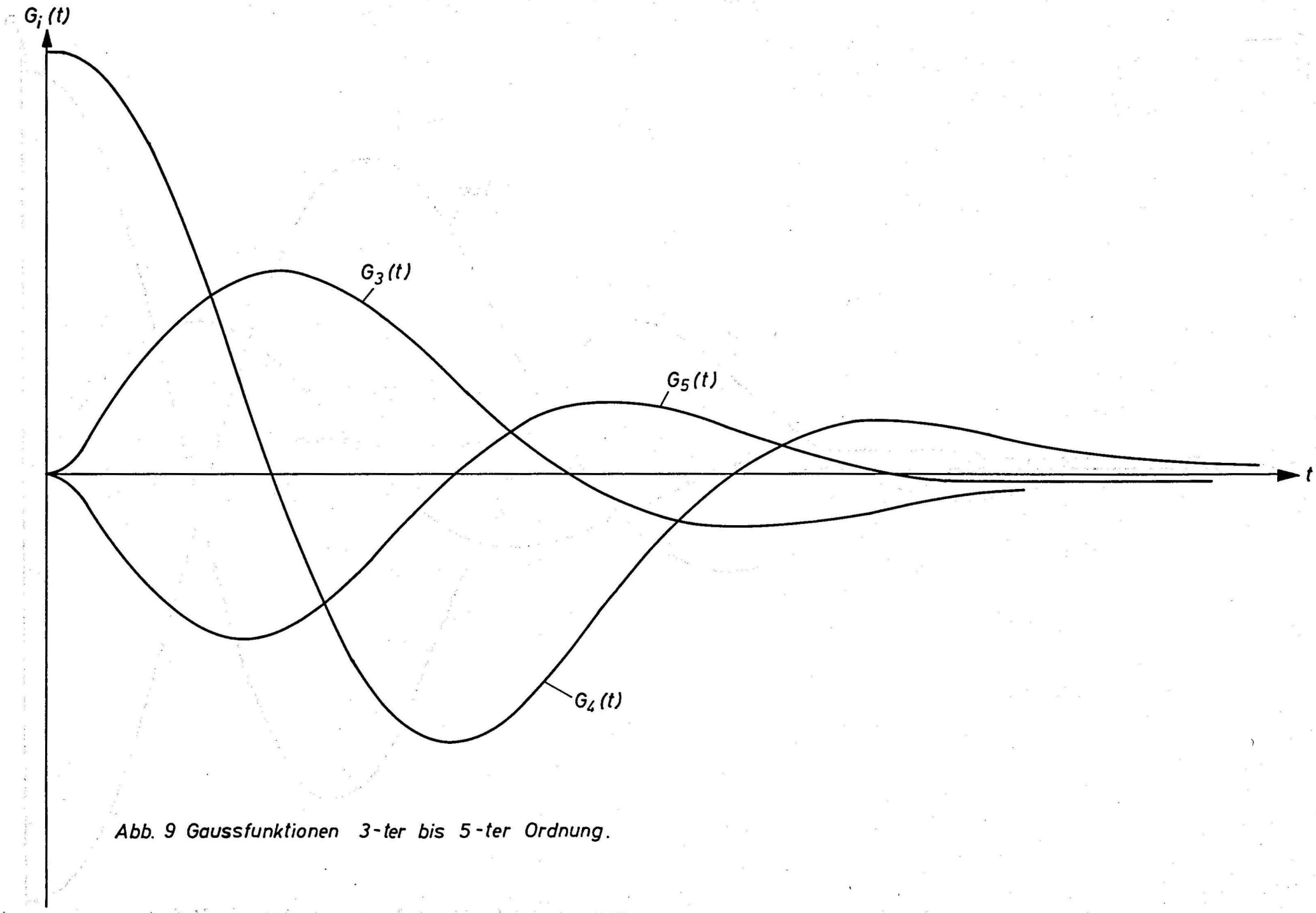

Abb. 9 Gaussfunktionen 3-ter bis 5-ter Ordnung.

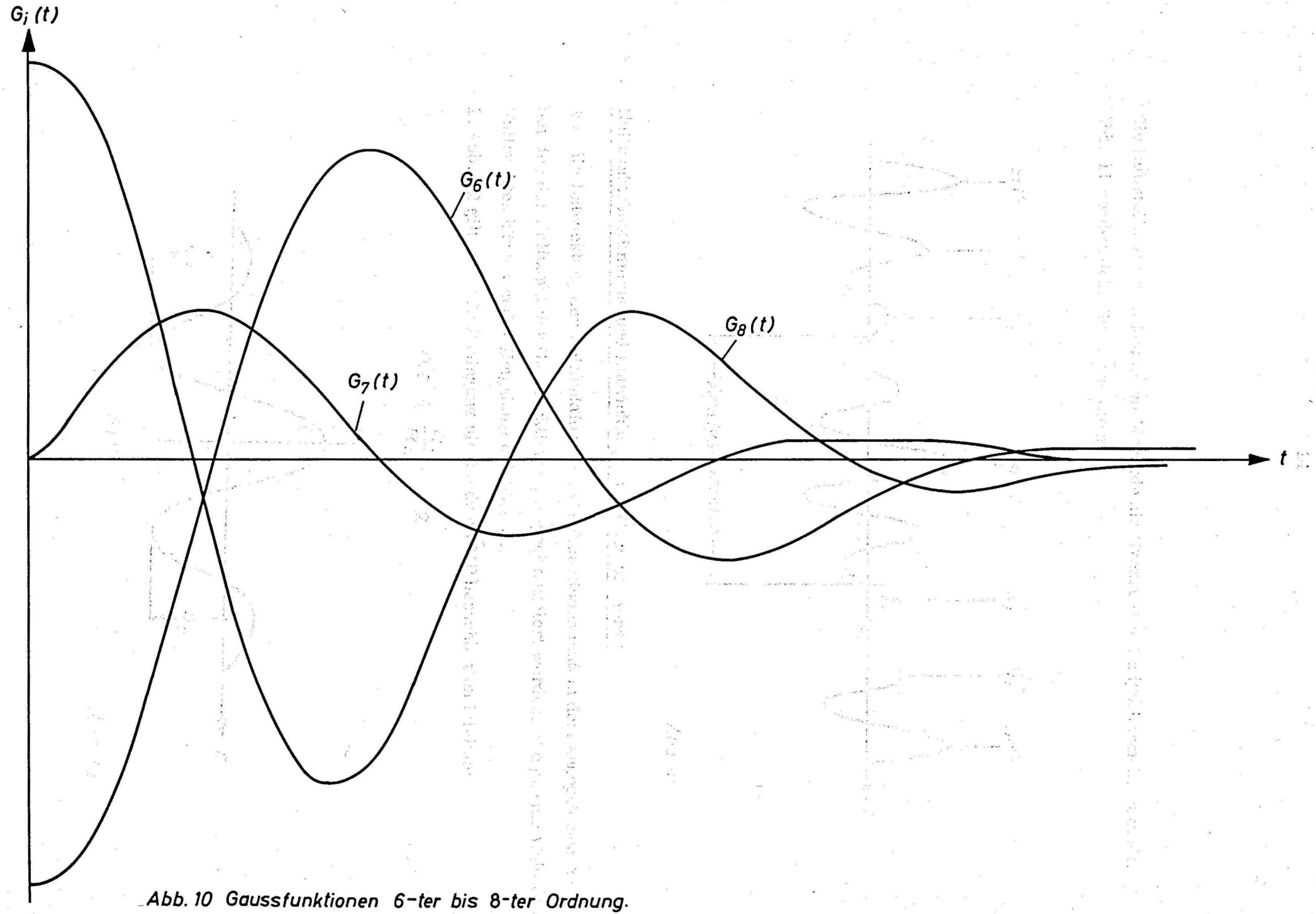

Abb. 10 Gaussfunktionen 6-ter bis 8-ter Ordnung.

Die Nullstelle vor dem Maximum M wurde als Grenze des Analysebereichs gewählt. Abb. 11 verdeutlicht die Segmentierung.

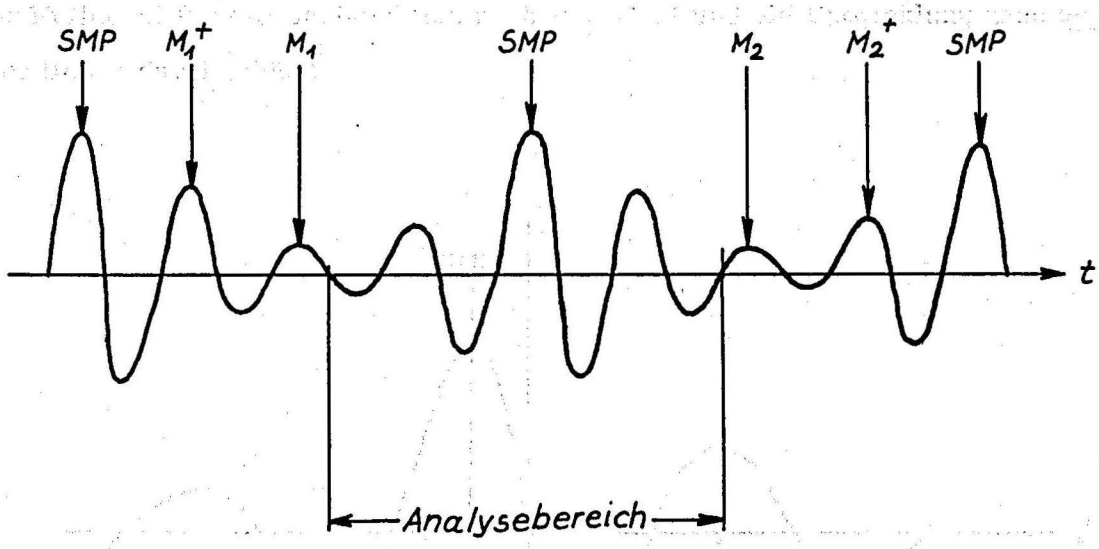

Abb. 11

3.3 Nullpunktsbestimmung und Entwicklung nach Gaußfunktionen

Für die Entwicklung der Zeitfunktion nach Gaußfunktionen muß in dem angegebenen Analysebereich der Nullpunkt der Gaußfunktionen geeignet definiert werden. Die genaue Wahl des Nullpunktes hängt von der Umgebung des SMP ab.

In Abb. 12 gilt folgende Bedingung für die dem SMP benachbarten Minima:

$$0.8 \leq \frac{a_1}{a_2} \leq 1.2$$

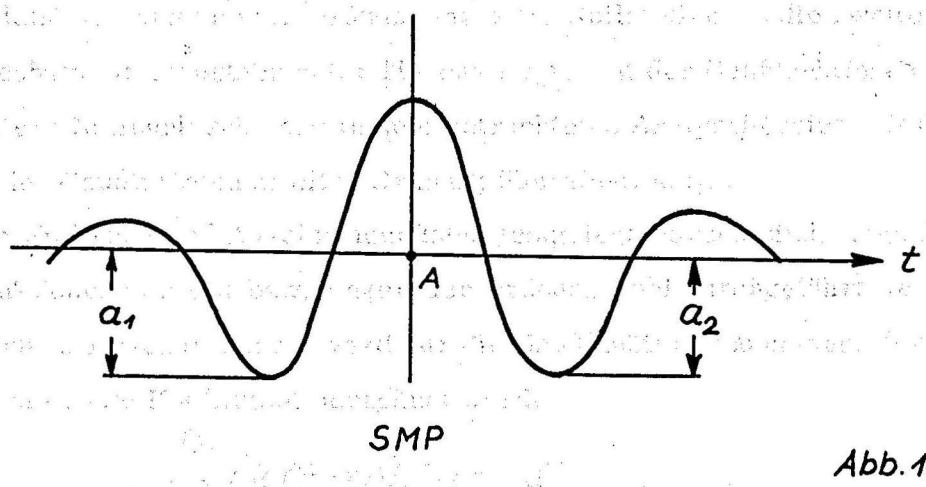

Abb. 12

In dem Fall wird der Nullpunkt zweckmäßigerweise nach A gelegt, d.h. an die Stelle des SMP und die Entwicklung nach geraden Gaußfunktionen durchgeführt.
Abb. 13 zeigt den Fall, in dem die oben erwähnte Bedingung nicht erfüllt ist. Dann wird der Nullpunkt für die Gaußfunktion nach B verlegt und die Entwicklung nach ungeraden Funktionen durchgeführt.

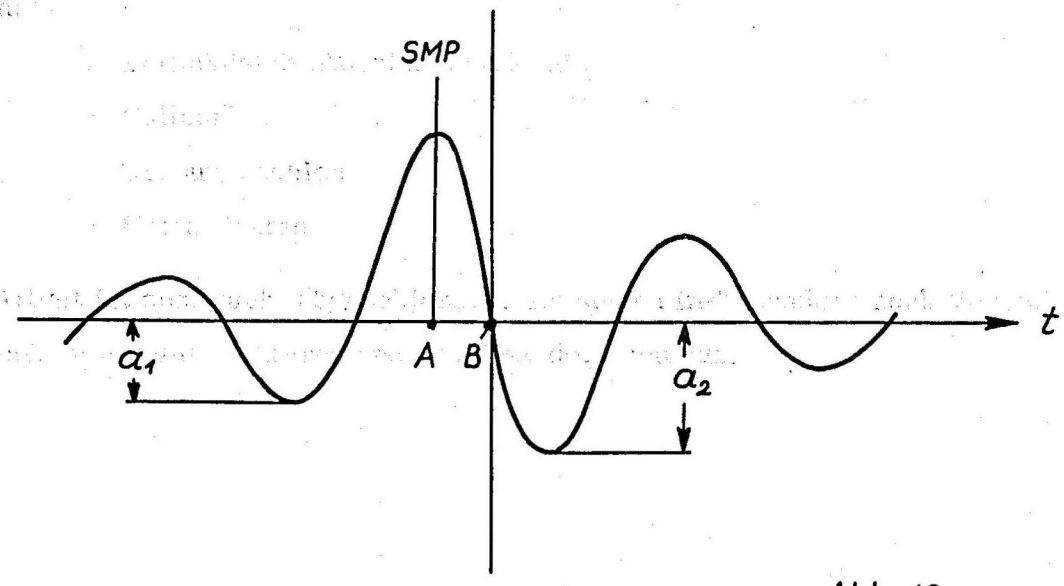

Abb. 13

Die Praxis hat gezeigt, daß die Approximation besser ist, wenn die Gaußfunktionen rascher abklingen. Das läßt sich durch die Wahl eines günstigen Abszissenmaßstabes erreichen. Die Gaußfunktion niedrigster Ordnung mit zwei Nullstellen ist die zweiter Ordnung. Es wurde deshalb ein Maßstabsfaktor für das Argument der Gaußfunktionen so bestimmt, daß der mittlere Nullpunktsabstand in dem betrachteten Analysebereich mit dem Nullpunktsabstand der Gaußfunktion zweiter Ordnung übereinstimmt.
Nachdem Nullpunkt und Abszissenmaßstab festgelegt worden sind, kann die Entwicklung nach Funktionen gerader bzw. ungerader Ordnungszahl durchgeführt werden. Als Ergebnis des ersten Analysedurchgangs wird nur die Gaußfunktion genommen, für die der dem Betrage nach größte Koeffizient berechnet wurde.

$$A_i = \frac{\int_{t_1}^{t_2} f[k(t-t_i)]G_i(t)\,dt}{\int_{t_1}^{t_2} G_i^2(t)\,dt} \qquad (12)$$

Für den zweiten Analysedurchgang wird der Zeitverlauf der im ersten Durchgang berechneten Gaußfunktion von der zu analysierenden Sprach-Zeitfunktion subtrahiert. Der Analysebereich und der Abszissenmaßstab bleiben erhalten, aber es wird ein neuer Nullpunkt gesucht.

Die Analyse wird nach dem oben beschriebenen Verfahren insgesamt dreimal durchgeführt, ehe zum angrenzenden Analysebereich übergegangen wird.

Das Ergebnis der Analyse für einen Analysebereich, d.h. für eine Pitchperiode sind 10 Daten:

1 Zeitmaßstab (Abszissenmaßstab)

3 Nullpunkte

3 Ordnungszahlen

3 Koeffizienten

Die Arbeit ist noch nicht abgeschlossen. Es stehen insbesondere noch Vergleiche mit den Sprachkompressionsfaktoren anderer Vocoderarten aus.

LITERATUR

1. FLANAGAN — Speech Analysis Synthesis and Perception. Springer-Verlag, Berlin, Heidelberg, New York, 1965.

2. L. R. RABINER — Digital-Formant-Synthesizer for Speech-Synthesis Studies. The Journal of the Acoustical Society of America, Vol. 43, No. 4, 1968

3. J. A. HOWARD, R. C. WOOD — Hybrid Simulation of Speech Waveforms Utilizing a Gaussian Wave Function Representation. Simulation, Sept. 1968

4. D. R. REDDY — Pitch Period Determination of Speech Sounds. Communications of the ACM, Vol. 10, No. 6, June 1967